Projet de TP : Classification et Régression

Dr. H.MOULAI

16 mars 2025

Objectif du Projet

L'objectif de ce projet est de permettre aux étudiants de mettre en pratique leurs connaissances en machine learning en travaillant sur deux problèmes distincts : un problème de **classification** et un problème de **régression**. Les étudiants devront choisir deux datasets adaptés, justifier leurs choix, et appliquer toutes les étapes nécessaires pour prétraiter les données, entraîner des modèles, et évaluer leurs performances.

1 Étapes du Projet

1.1 Choix des Datasets

Les étudiants doivent choisir :
- Un dataset pour un problème de **classification** (e.g., prédiction de classes binaires ou multiclasses).
- Un dataset pour un problème de **régression** (e.g., prédiction de valeurs continues).

Critères de choix :
- Les datasets doivent être suffisamment riches (nombre d'échantillons et de caractéristiques).
- Les datasets doivent être accessibles (via des sources comme Kaggle, UCI Machine Learning Repository, ou des APIs publiques).
- Justifier le choix des datasets en expliquant leur pertinence pour les problèmes de classification et de régression.

Exemples de datasets :
- Classification : Iris, MNIST, Titanic, Breast Cancer Wisconsin.
- Régression : Boston Housing, California Housing, Diabetes, Wine Quality.

1.2 Prétraitement des Données

Les étudiants doivent appliquer les étapes suivantes pour chaque dataset :

1. **Nettoyage des données** :

- Gestion des valeurs manquantes (imputation ou suppression).
- Suppression des doublons.

2. **Exploration des données** :
   - Analyse statistique (moyenne, médiane, écart-type, etc.).
   - Visualisation des données (histogrammes, boxplots, heatmaps de corrélation).
3. **Transformation des données** :
   - Encodage des variables catégorielles (One-Hot Encoding, Label Encoding).
   - Normalisation ou standardisation des données.
4. **Séparation des données** :
   - Diviser les données en ensembles d'entraînement et de test (e.g., 80/20 ou 70/30).

1.3 Modélisation

Pour chaque problème (classification et régression), les étudiants doivent :

1. **Choix des modèles** :
   - Classification : Modèles comme la régression logistique, les arbres de décision, les forêts aléatoires, ou SVM.
   - Régression : Modèles comme la régression linéaire, les arbres de décision, les forêts aléatoires, ou le gradient boosting.
2. **Entraînement des modèles** :
   - Entraîner les modèles sur l'ensemble d'entraînement.
3. **Optimisation** :
   - Utiliser la validation croisée (cross-validation) pour ajuster les hyperparamètres.
   - Appliquer des techniques comme GridSearchCV ou RandomizedSearchCV.

1.4 Évaluation des Modèles

Les étudiants doivent évaluer les performances des modèles en utilisant des métriques adaptées :

- **Classification** :
  - Métriques : Accuracy, Precision, Recall, F1-Score, Matrice de confusion.
  - Visualisation : Courbe ROC, AUC.
- **Régression** :
  - Métriques : MSE (Mean Squared Error), RMSE, MAE, R^2.
  - Visualisation : Graphiques des résidus, comparaison des prédictions aux valeurs réelles.

1.5 Analyse des Résultats

— Comparer les performances des différents modèles.
— Identifier les modèles les plus performants et expliquer pourquoi.
— Discuter des limites des modèles et des éventuelles améliorations possibles.

1.6 Rapport et Présentation

Les étudiants doivent rédiger un rapport détaillé et préparer une présentation orale. Le rapport doit inclure :

1. **Introduction** :
   — Contexte du projet et objectifs.
2. **Choix des datasets** :
   — Description des datasets et justification des choix.
3. **Prétraitement des données** :
   — Détails des étapes de nettoyage et de transformation.
4. **Modélisation** :
   — Description des modèles utilisés et des techniques d'optimisation.
5. **Résultats** :
   — Analyse des performances des modèles.
6. **Conclusion** :
   — Synthèse des résultats et perspectives d'amélioration.

2 Critères d'Évaluation

— **Choix des datasets** (10%) :
  — Pertinence et justification.
— **Prétraitement des données** (20%) :
  — Qualité du nettoyage et de la transformation.
— **Modélisation** (30%) :
  — Diversité des modèles testés et optimisation.
— **Évaluation des performances** (20%) :
  — Utilisation appropriée des métriques et analyse des résultats.
— **Rapport et présentation** (20%) :
  — Clarté, structure, et qualité de la présentation.